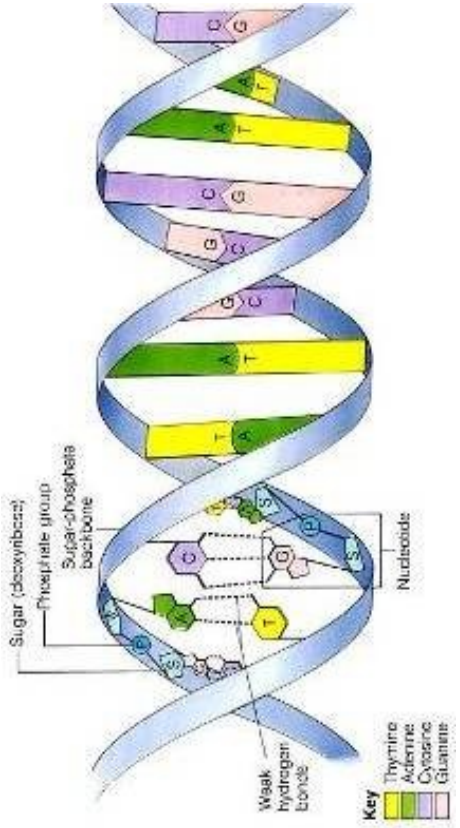
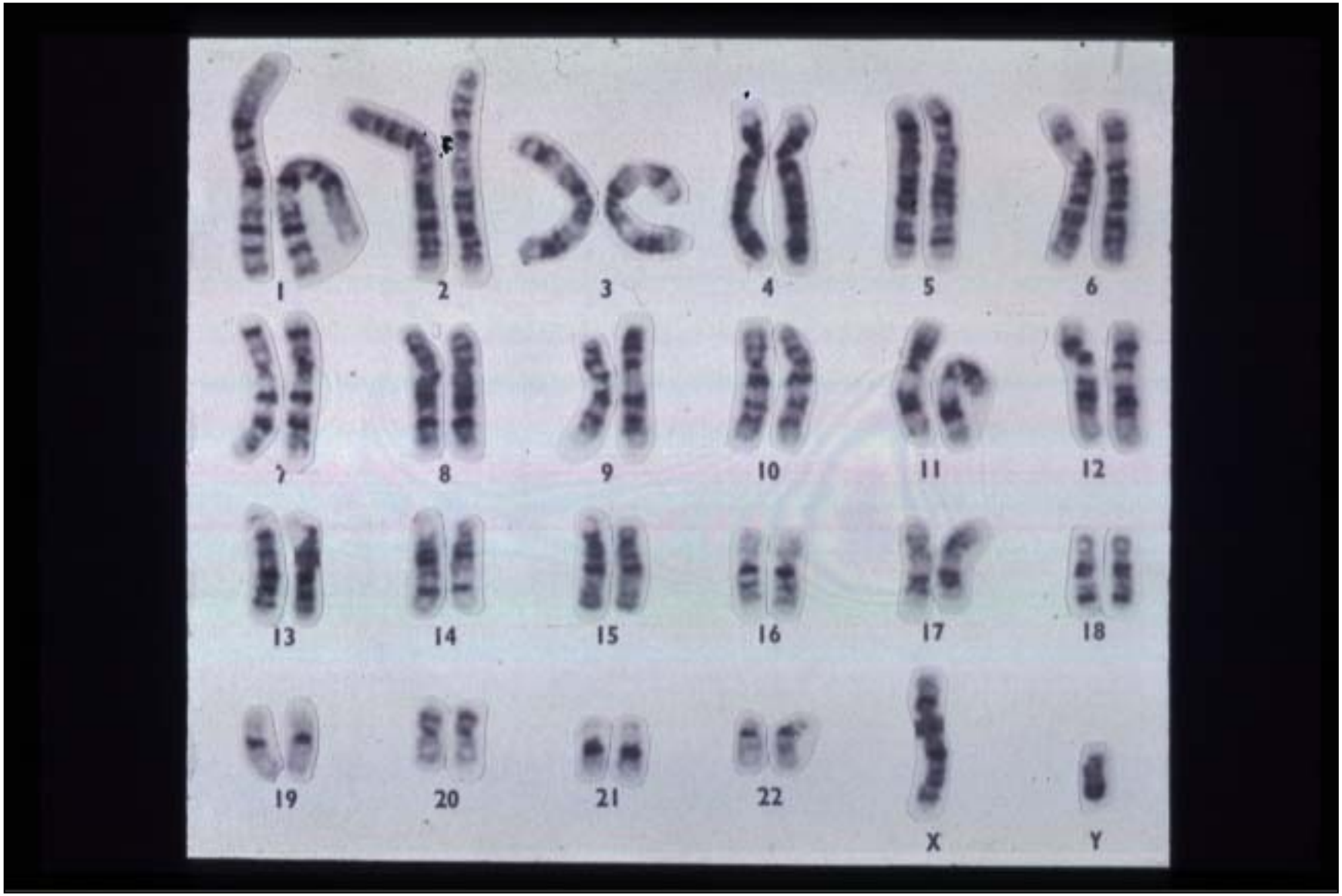
Gene mapping
I. Model-based (parametric)



Gene mapping – why?

- Objective is to find genes linked to traits of interest e.g. disease
- Identify gene and its function
- Understand how mutation causes disease
- Develop new therapeutic methods or drugs
- DNA testing and estimation of genetic risk

Where is the gene? 3000 Mb 24 chromosomes



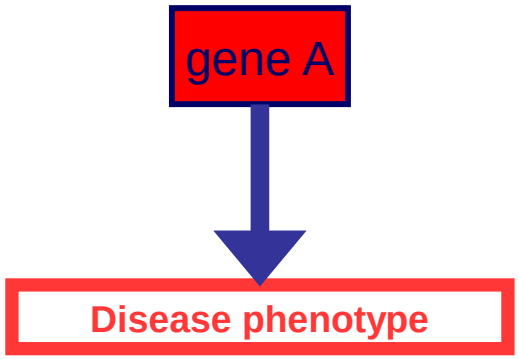
Is it genetic?

- Record information about affected individuals and their family relationships (pedigree) – clustering in families
- Relative risk / twin studies
- Segregation analysis
- Heritability analysis
- Single gene (Mendelian) or complex?

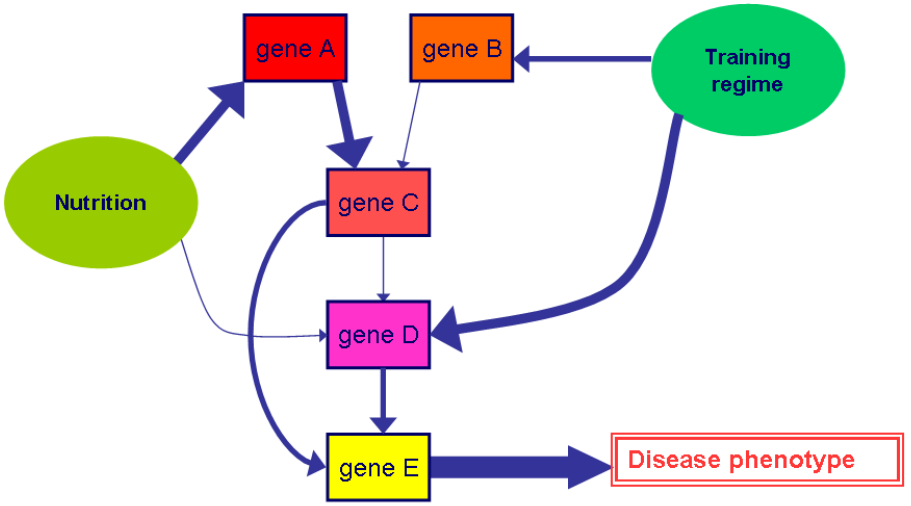
  (Is it showing Mendelian Inheritance?)

Genetic architecture

Single gene or Mendelian



Complex disease



Parametric linkage

Non-parametric linkage
Association

Single Gene Disorders

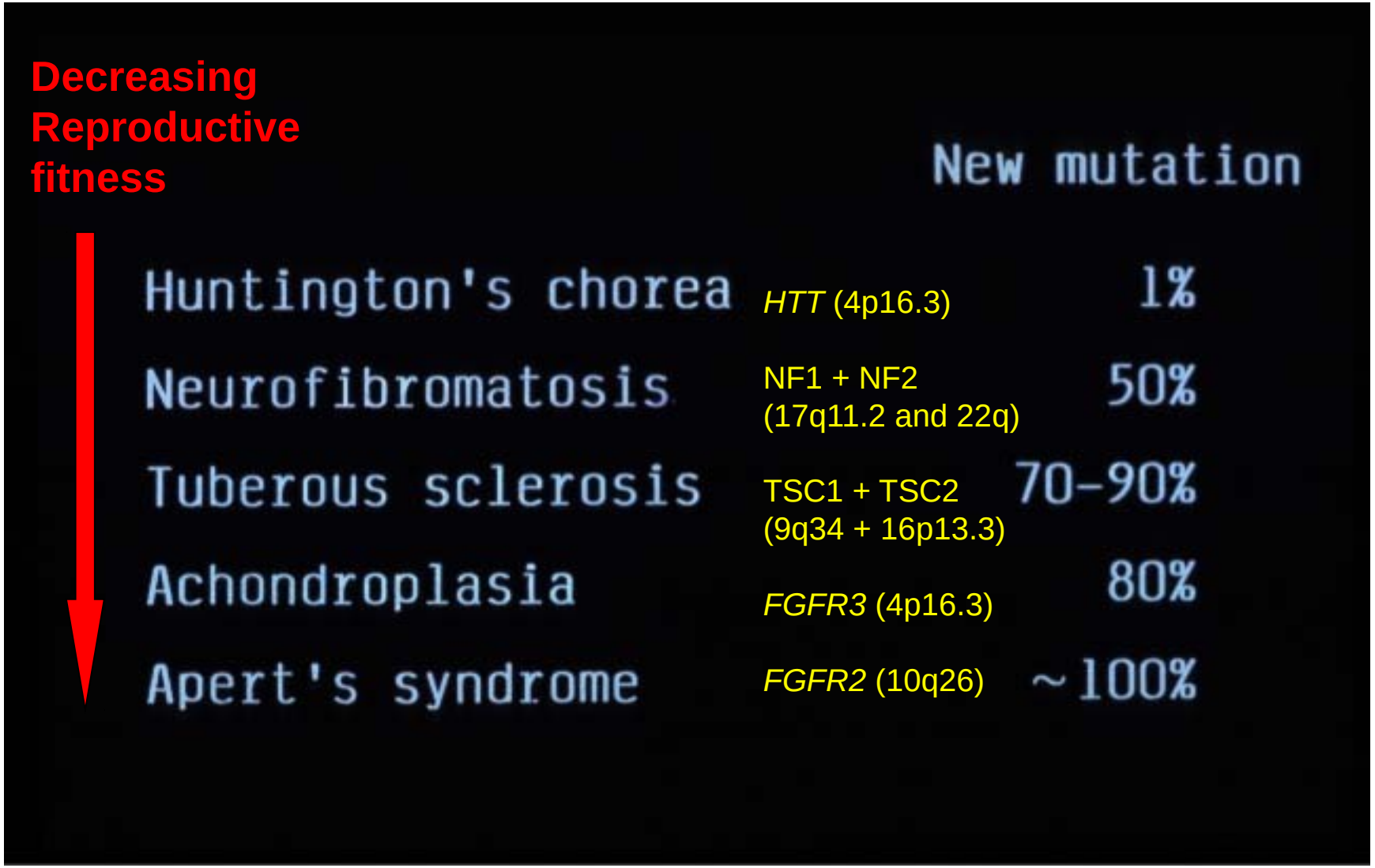
6237 single gene disorders

Collective genetic burden – 1 : 200 births

Not all mapped – available families

Not all are inherited – eg autosomal dominants with high penetrance and low reproductive fitness – new mutation

Single Gene Disorders



Single Gene Disorders

Reduced Reproductive Fitness

Reduced survival to reproductive age.

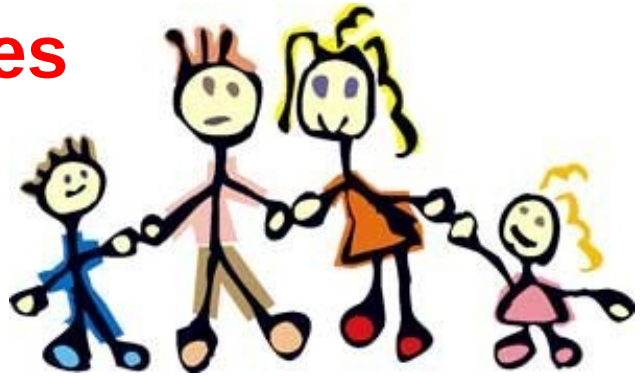
- Increased foetal loss
- Increased perinatal mortality
(including obstetric complications)
- Increased childhood/adolescent mortality

Reduced chance of finding a partner.

Impaired fertility.

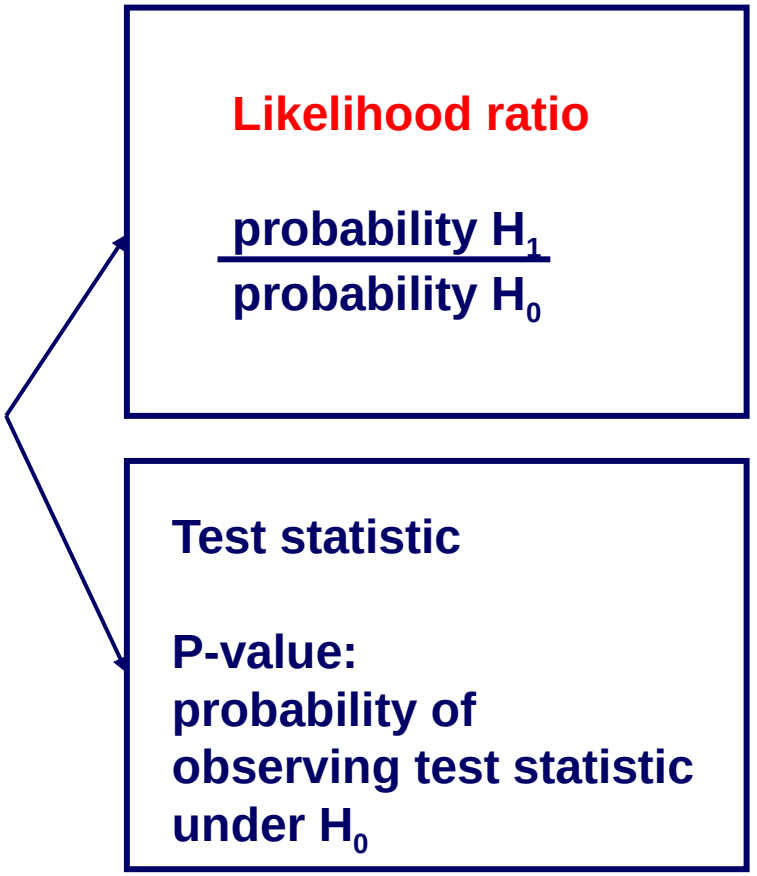
Gene mapping by linkage

- Collect DNA samples from individuals belonging to families
- Genotype collection of genetic markers along the genome
- Calculate an appropriate linkage statistic for each marker
- Identify regions where statistic differs significantly from expectation when marker is not linked to trait



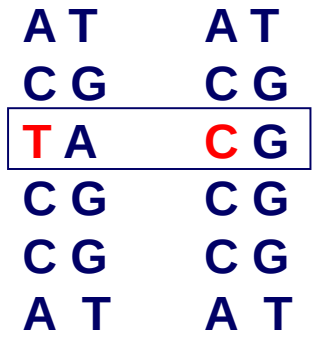
Hypothesis testing

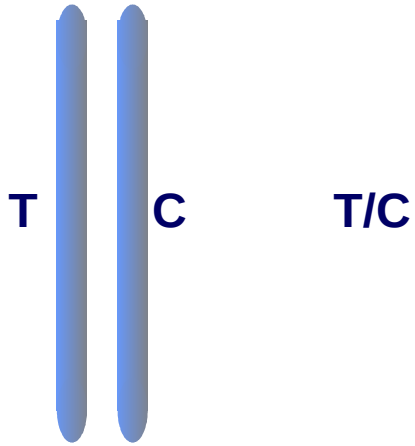
- Null hypothesis (H_0)
- Alternative hypothesis (H_1) Linked
- Deviation from the null hypothesis indicates linkage



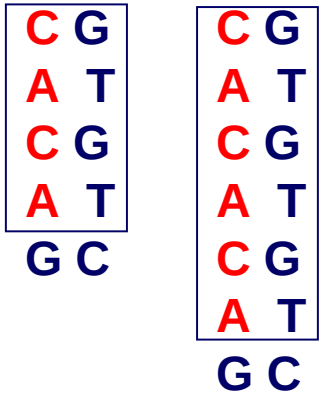
Genetic markers

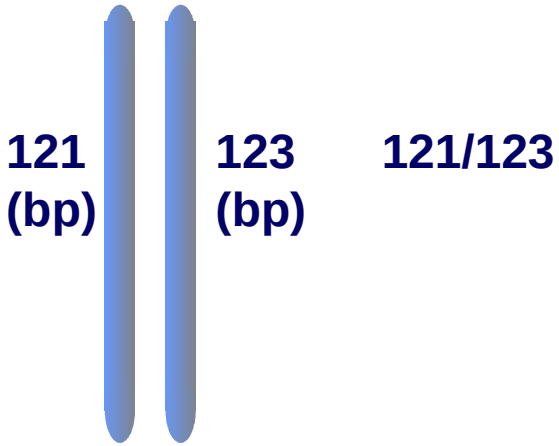
Single Nucleotide Polymorphism (SNP)



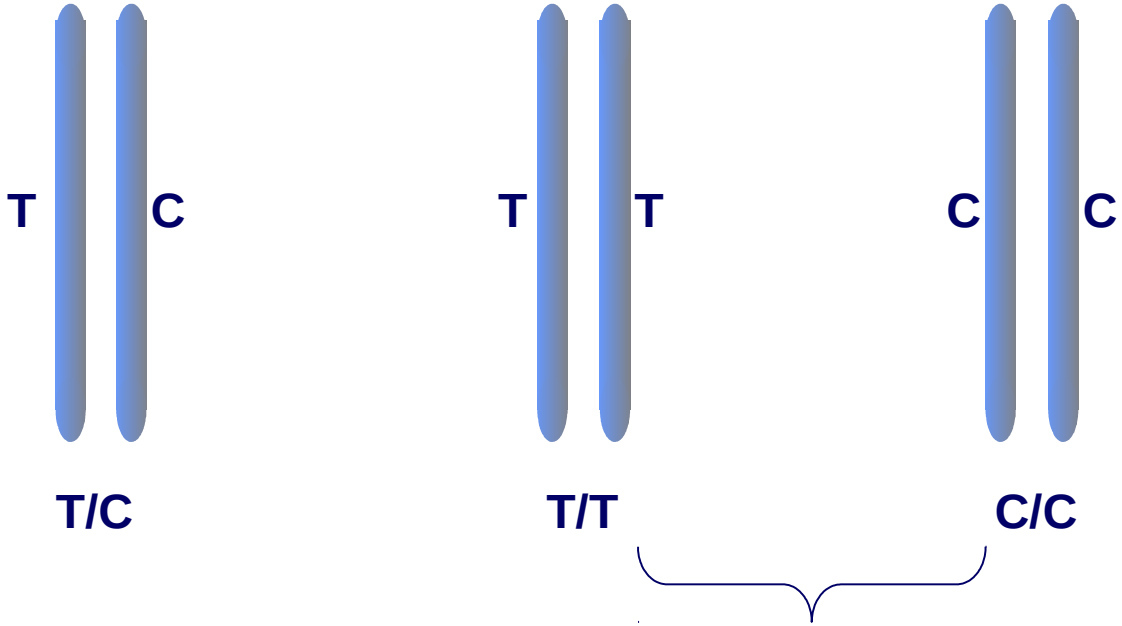


Microsatellite





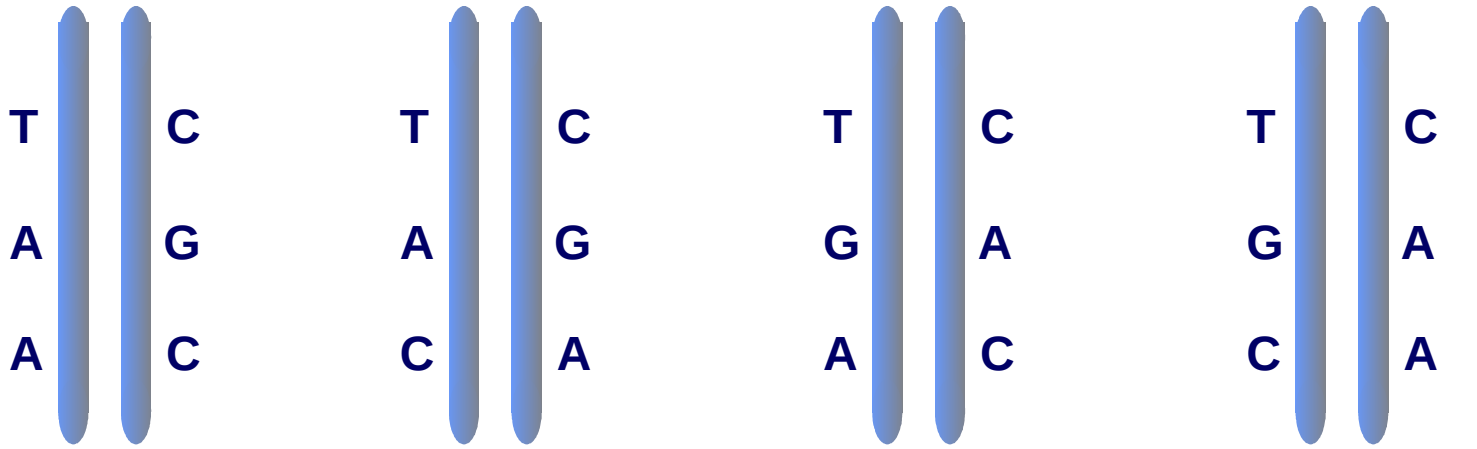
Genotypes



Frequency	**2pq**	**p^2**	**q^2**
Penetrance (probability of phenotype given a defined genotype)	**f_1**	**f_0**	**f_2**

Haplotype phase

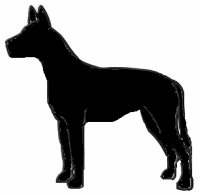
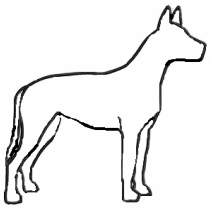
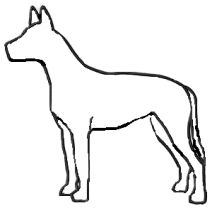
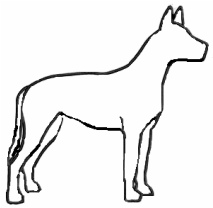
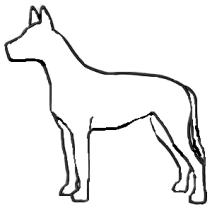
Genotype **T/C A/G C/A**



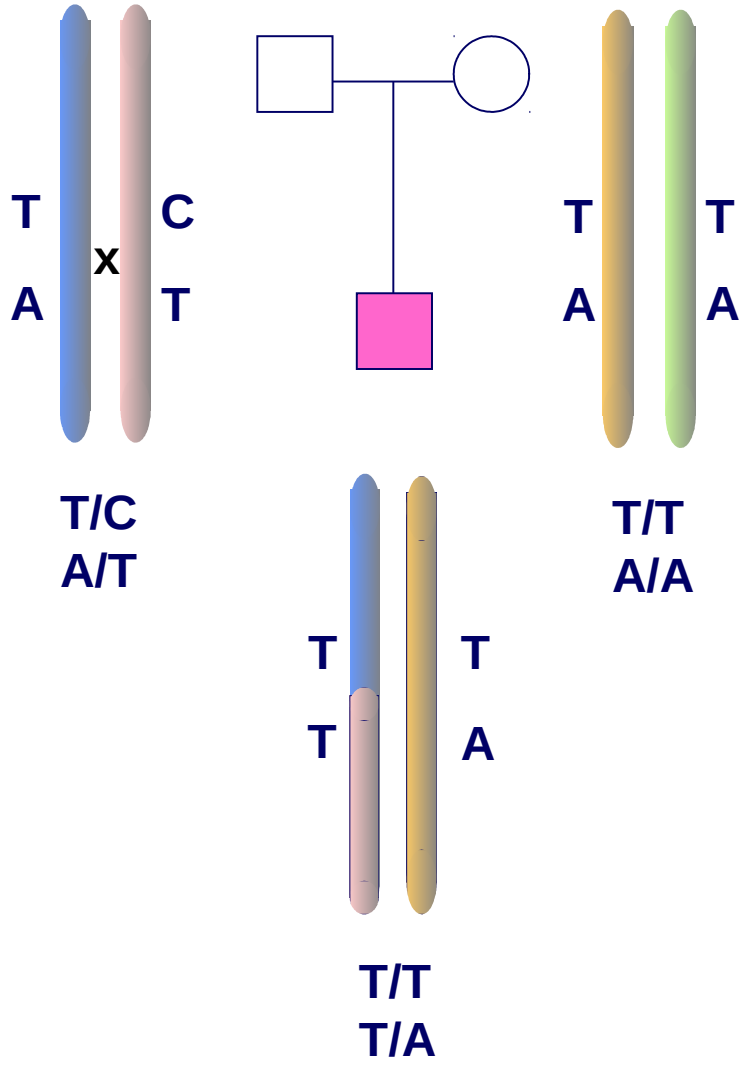
TAA / CGC **TAC / CGA** **TGA / CAC** **TGC / CAA**

Haplotype

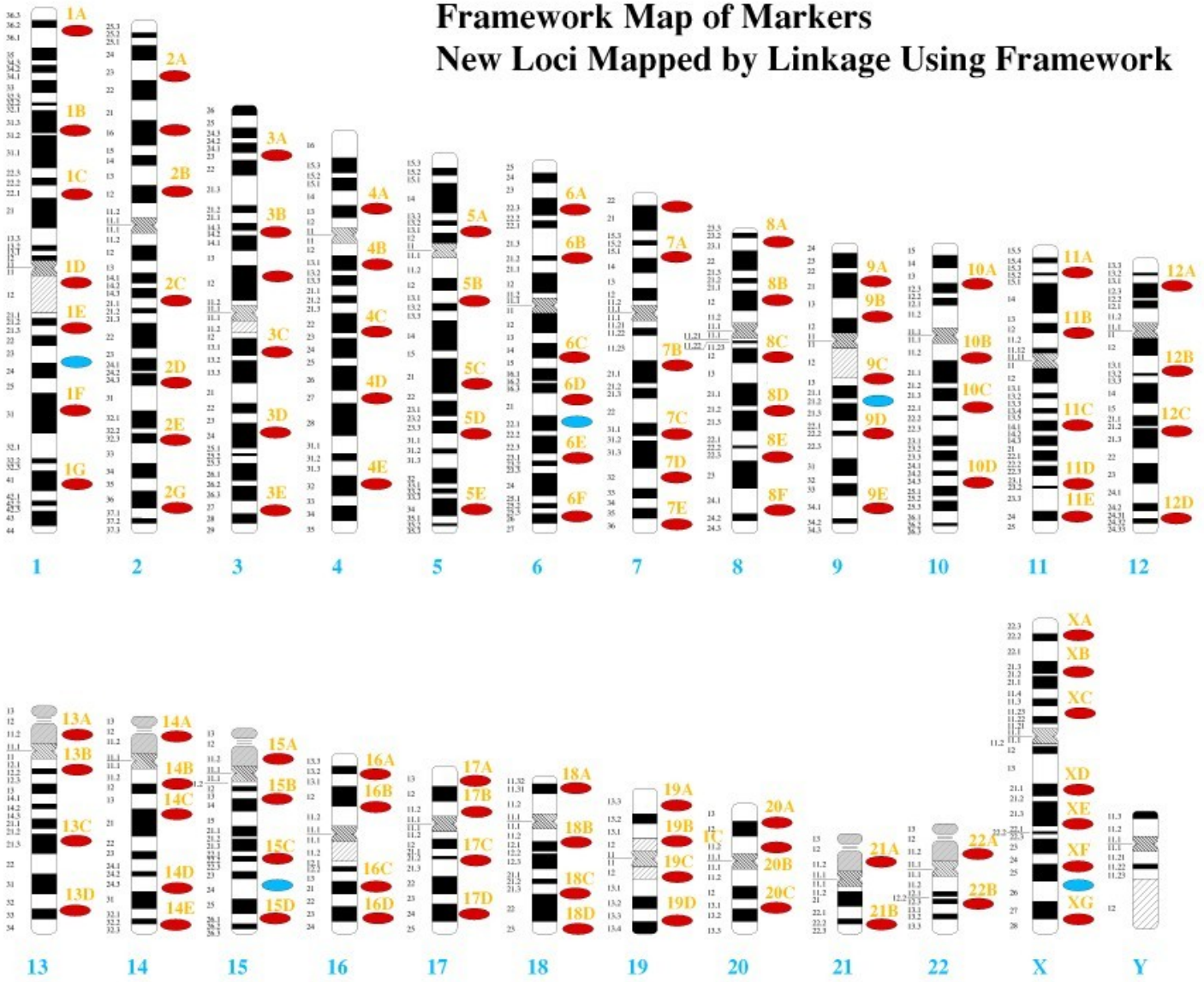
Mode of inheritance

Genotype	Phenotype		
T/T			
T/C			
C/C			
Single gene or Mendelian trait	Co-dominant	Recessive	Dominant

Informative markers

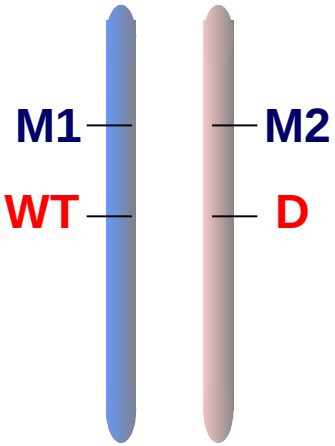


Principles of linkage



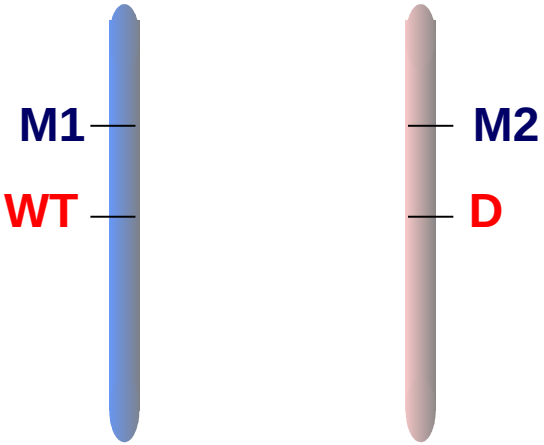
Principles of linkage

Complete linkage



Zero recombination

Meiosis

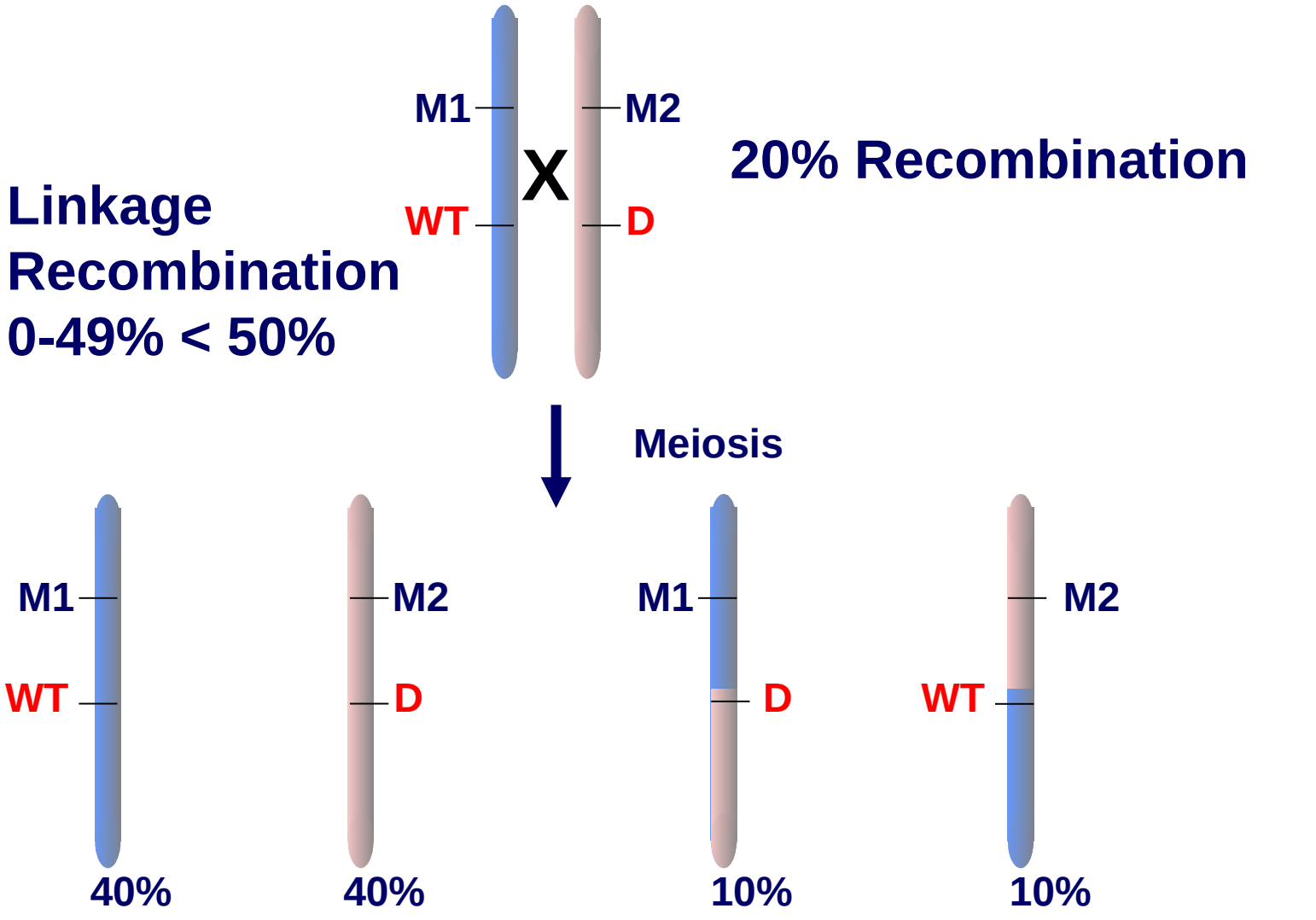


100% non-recombinant

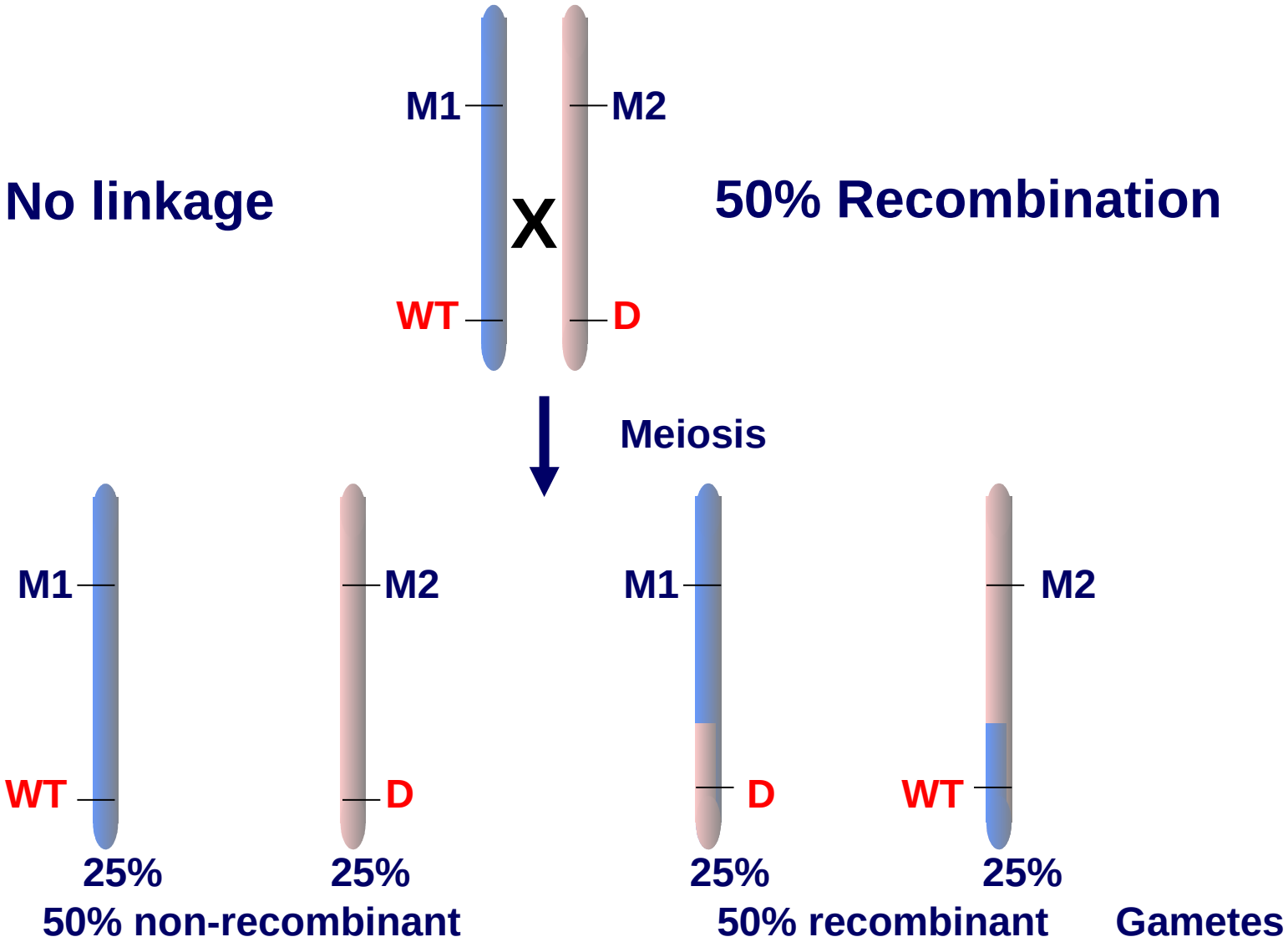
0% recombinant

Gametes

Principles of linkage

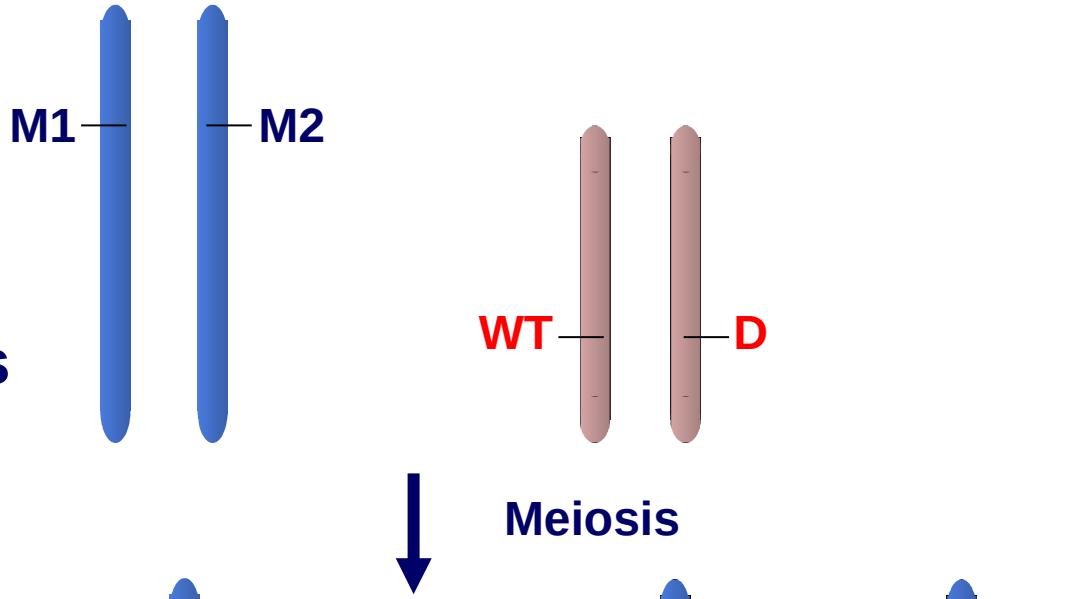


Principles of linkage



Principles of linkage

Disease and marker on different chromosomes



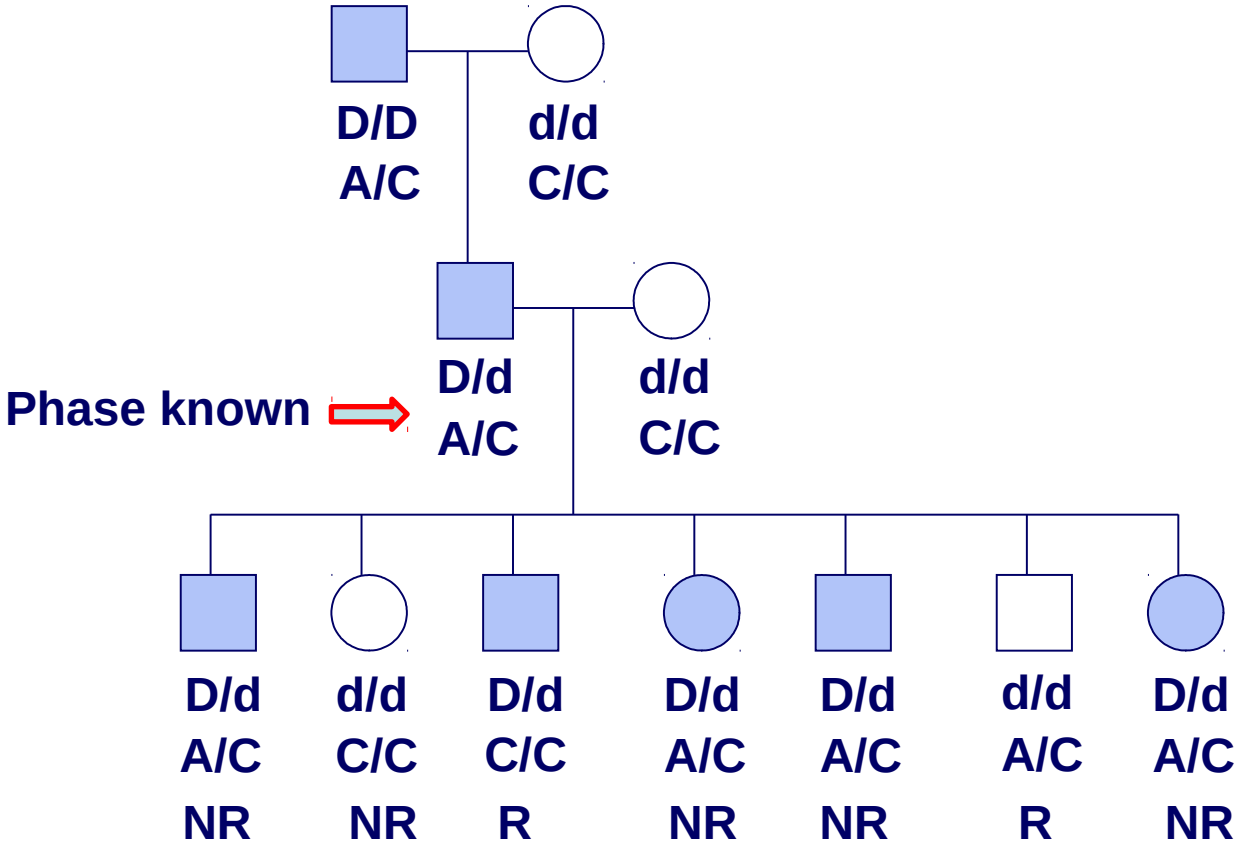
Principles of linkage

	Loci on same chromosome			Loci on different chromosomes
	Close	Nearby	Distant	
Crossing over between loci	rare	some	frequent	frequent
Linkage	present	present	absent	absent
Recombination Fraction (θ)	0%	1 - 49%	50%	50%

Parametric linkage analysis

- Tests for co-segregation between a trait and marker by conceptualizing the trait phenotype as a hidden genotype
- Requires specification of penetrances (probability of disease, given genotype status). Most monogenic disorders 100%
- Requires specification of trait gene frequencies
- Gene frequencies and genotype penetrances are parameters

Simple disease model: single dominant gene



Calculating a LOD score (Z)

$LOD = \log_{10}(Likelihood_{H1}/Likelihood_{H0})$

H0: recombination rate (θ) = 0.5
no linkage

H1: $0 <= \theta < 0.5$
linkage

Calculating a LOD score

For the example family:

H0: if $\theta = 0.5$, likelihood = $(0.5)^7$

H1: likelihood of 5 non-recombinants and 2 recombinants
= $(1 - \theta)^5 \theta^2$

1 = 100% linkage since recombination θ will be zero

$$\text{LOD} = \log_{10} \left(\frac{(1 - \theta)^5 \, \theta^2}{(0.5)^7} \right)$$

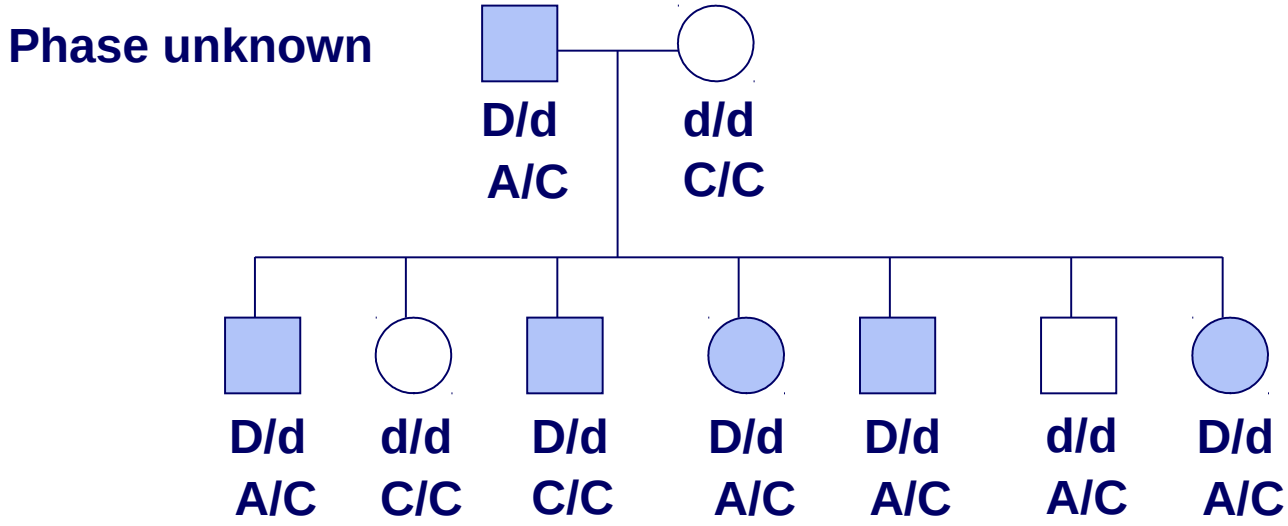
Estimating recombination rate

Recombination fraction (θ)

	0.01	0.05	0.1	0.2	0.3	0.4
LOD	-1.92	-0.61	-0.12	0.23	0.29	0.20

Maximum LOD score

Simple disease model: single dominant gene



D/d A/C	Phase 1	NR	NR	R	NR	NR	R	NR
D/d C/A	Phase 2	R	R	NR	R	R	NR	R

Calculating a LOD score

Phase 1 Phase 2

$$LOD = \log_{10}\left[\frac{0.5\,((1-\theta)^5\,\theta^2)+0.5((1-\theta)^2\,\theta^5)}{(0.5)^7}\right]$$

Recombination fraction (θ)

	0.01	0.05	0.1	0.2	0.3	0.4
LOD	-2.21	-0.87	-0.35	0.1	0.29	0.38

Maximum LOD score

Significance

Lod score	Interpretation for autosomal loci
≥ + 3	Linkage established (~ 95% confidence)
< -2	Linkage excluded at that recombination fraction
-2 → +3	Inconclusive, need more data (more families)

Overall probability in a set of families is the product of the probabilities in individual families. Since Z (lod score) is a log then they can be simply added

Significance

Hypothesis	Linked Recombination = θ	Not linked Recombination = 50%
Prior probability	0.02	0.98
Conditional probability (LOD score>3)	1000	1
Joint probability prior x conditional (0.02 x 1000)	20	~1 (~p-value 0.05)

Gene mapping by linkage

1. Ascertain families and obtain DNA samples
2. Type for set of genetic markers across genome
3. For each marker, calculate LOD scores for each family at different values of recombination fraction θ

   (say for θ = 0, 0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4)
4. For each value of θ add LOD scores across all families in order to get sufficient meioses to measure recombination fraction and determine linkage.
5. Positive LOD scores suggest linkage
6. Highest LOD score indicates most likely value of θ
7. Negative LOD scores suggest no linkage

Types of linkage analysis

- **Two point analysis**

  θ is maximized between two loci,

  the assumed phenotypic trait or disease locus and a polymorphic marker

- **Multipoint analysis**

  Several polymorphic marker loci, of known order (on the genetic map)

  Trait locus shifted along genetic map until LOD score is maximized

What can go wrong?

- Incorrect determination of phenotype (disease status)
- Wrong or incomplete genetic model (e.g. locus heterogeneity)
- Non-paternity or pedigree errors
- Laboratory errors
  - samples muddled
  - genotype error
  - data recording error

Recombination Fraction: Genetic Distance: Physical Distance

- 1% recombination = 1cM
- 49 Chiasma / male genome each yielding

  50% recombinants = ~ 2500cM

  Female genome more chiasma = ~ 4500cM
- Genome = 3000Mb Male 1cM = 1.13Mb

  Female 1cM = 0.67Mb

  Sex average = ~ 0.9Mb

Recombination Fraction: Genetic Distance: Physical Distance

- Genome distribution of chiasma not random
- Interference between adjacent cross-overs
- Relation between recombination fraction, genetic and physical distance varies in different genomic regions – more telomeric than centromeric recombination
- Development of mapping functions to relate recombination fraction to genetic distance

eg Kosambi's Function $\omega=0.25\ln[(1+2\theta)/(1-2\theta)]$

where ω = map distance and θ = Recomb. fraction

Polygenic and Multifactorial Inheritance

Polygenic Inheritance = traits / diseases caused by the impact of many different genes each having a small individual effect on phenotype.

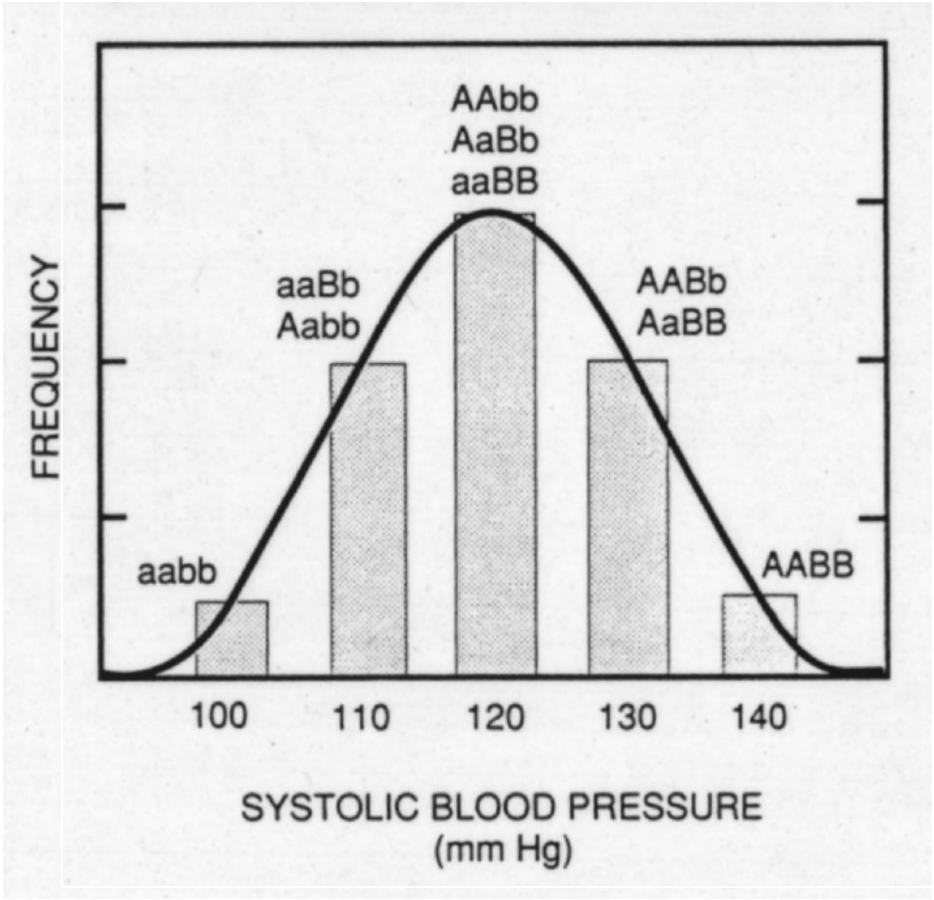
Multifactorial Inheritance =
The occurrence of the condition is dependent on the combined effect of :

1. **Several genes each exerting a small individual influence on phenotype.**

2. **Interplay of several environmental factors (each exerting a small influence) with multiple genes.**

This definition provides a model that is amenable to statistical analysis. In reality, the number of genes and environmental factors involved could be small.

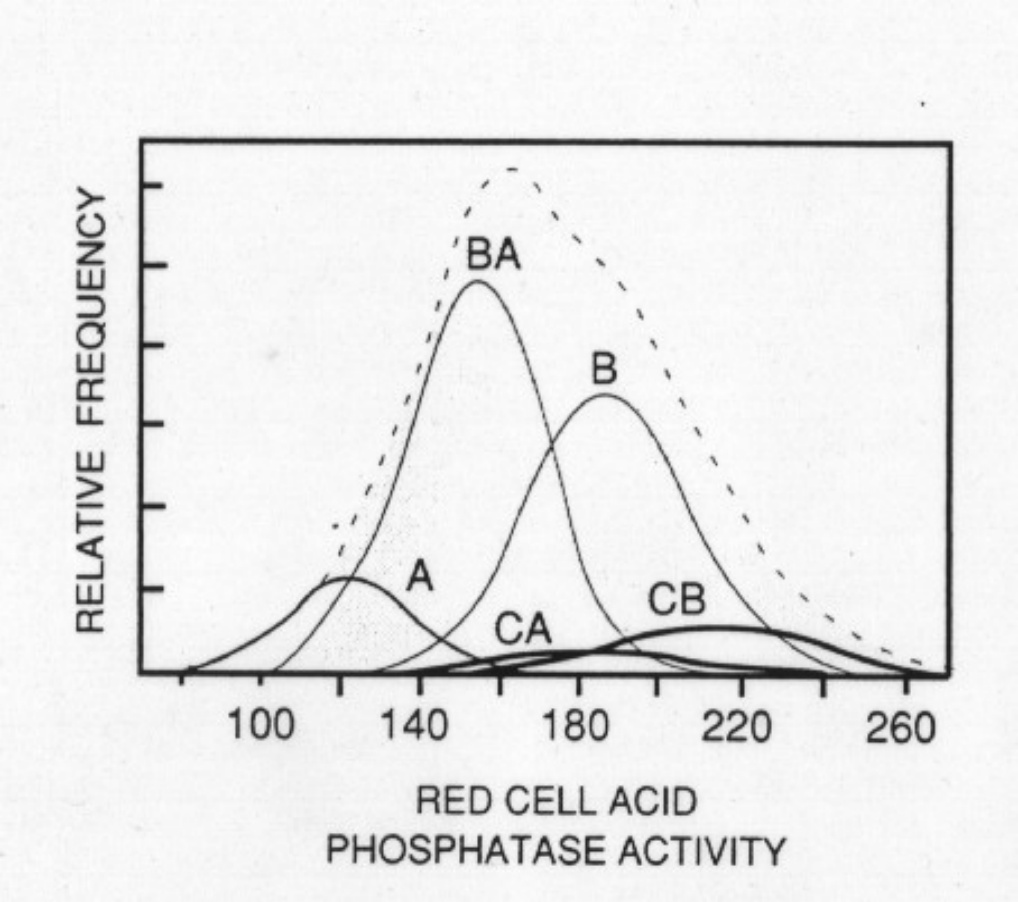
Polygenic / Multifactorial traits are quantitative and distributed continuously in the population approximating to a normal frequency distribution.



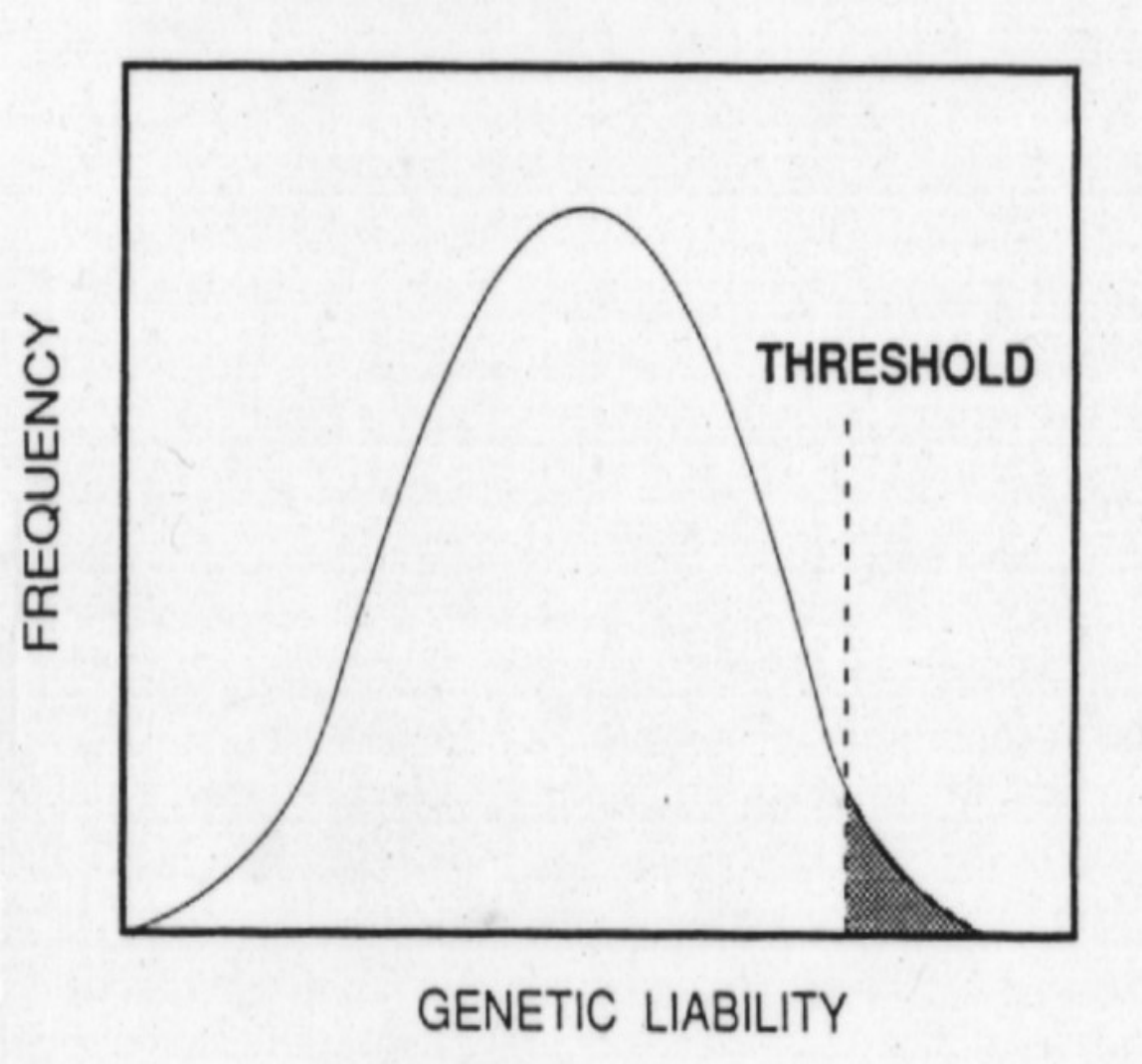
Hypothetical 2 locus / 2 allele system controlling systolic blood pressure. Normal distribution of genotypes. Increasing numbers of dominant alleles > higher systolic blood pressure.

In reality the unimodal distribution of a continuous phenotype in the population represents the summation of several discrete phenotypes created by combinations of alleles at the contributing loci.

eg Red cell acid phosphatase activity; Three alleles and six phenotypes. CC is rare.



Variant alleles of genes in the general population can confer a genetic predisposition to disease and appear as discontinuous disease traits. This reflects the threshold of genetic liability (combinations of predisposing alleles at different loci) required to lead to disease.



Linkage mapping of a complex trait

- Double-muscling in cattle (generalized muscular hypertrophy)
- First documented description in 1807
- Reported in several cattle breeds
- Segregation analysis in Belgian Blue cattle indicated a monogenic, autosomal recessive segregation pattern

Linkage mapping of a complex trait

Belgian Blue "double-muscled" cattle

Linkage mapping of a complex trait

Experimental cross

mh/mh

+/+

F1

+/mh

mh/mh

Backcross

108 calves

+/mh

mh/mh

Linkage mapping of a complex trait

- 213 microsatellite markers distributed over the 29 bovine autosomes were genotyped
- Linkage analysis carried out, model applied
  - monogenic, biallelic gene (*mh* and +)
  - full penetrance for *mh*/*mh*
  - zero penetrance for *mh*/+ and +/+
- Evidence for linkage on chromosome 2

Charlier *et al* (1995) Mammalian Genome 6, 788-792

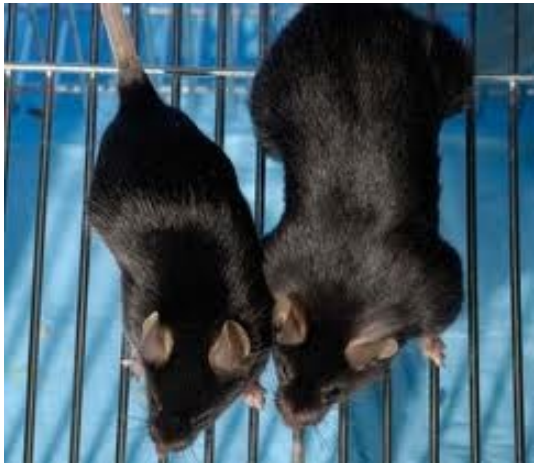
Linkage mapping of a complex trait

	Recombination fraction θ			
Marker	0	0.01	0.05	0.1
TGLA110	- ∞	-28.9	-14.6	-8.9
TGLA226	- ∞	-35.0	-17.5	-10.4
BM4440	- ∞	-35.4	-16.7	-9.3
ETH121	- ∞	-34.0	-14.6	-7.3
TGLA377	- ∞	-19.3	-6.8	-2.3
TGLA431	- ∞	-4.2	4.4	7.1
TGLA44	- ∞	15.8	15.4	14.3

Support interval of 12cM

Linkage mapping of a complex trait

- **Mouse model gave clue to candidate gene in the region identified**



- **Mutation identified as a deletion in the *myostatin* gene (Grobet *et al* 1997 Nature Genetics 17, 71-74)**

Linkage mapping of a complex trait

- Five independent mutations found in *myostatin* in different cattle breeds showing double muscling
- But some breeds are homozygous for the deletion and show no evidence of double muscling e.g. Highland Cattle
  1. myostatin making large contribution 2. other genes making smaller contribution. Highland may have another mutation that counteracts myostatin.
- Double muscling a true complex trait that involves several genes

Linkage mapping of a complex trait

- Mapping complex trait genes possible under the right circumstances
- BUT data structure needs to be suitable (clear phenotype and pedigree structure)
- Genes need to be acting in a near Mendelian fashion (large genetic contribution)
- Most of the time pretty difficult and other approaches need to be taken – non-parametric linkage or association where there are small contributions from many genes

Reading

Human Molecular Genetics 4
Strachan & Read, Garland Science 2011

Chapter 14 Genetic mapping of Mendelian characters

Chapter 15 Mapping and Identifying Genes Conferring Susceptibility to Complex Diseases

Generally several other chapters relevant to this Section of the course

Analysis of Human Genetic Linkage
Ott, John Hopkins University Press 1999

Linkage analysis software

MERLIN software
http://www.sph.umich.edu/csg/abecasis/merlin/index.html

Parametric linkage analysis - tutorial
http://www.sph.umich.edu/csg/abecasis/merlin/tour/parametric.html